RNA-Seq workflow

Phase 1: developing a workflow and preprocessing raw reads

Andrew Ndhlovu

Project SeaStore
Department of Botany and Zoology
Stellenbosch University
https://github.com/PiscatorX/RNA-Seq-devs

June 26, 2023

Overview

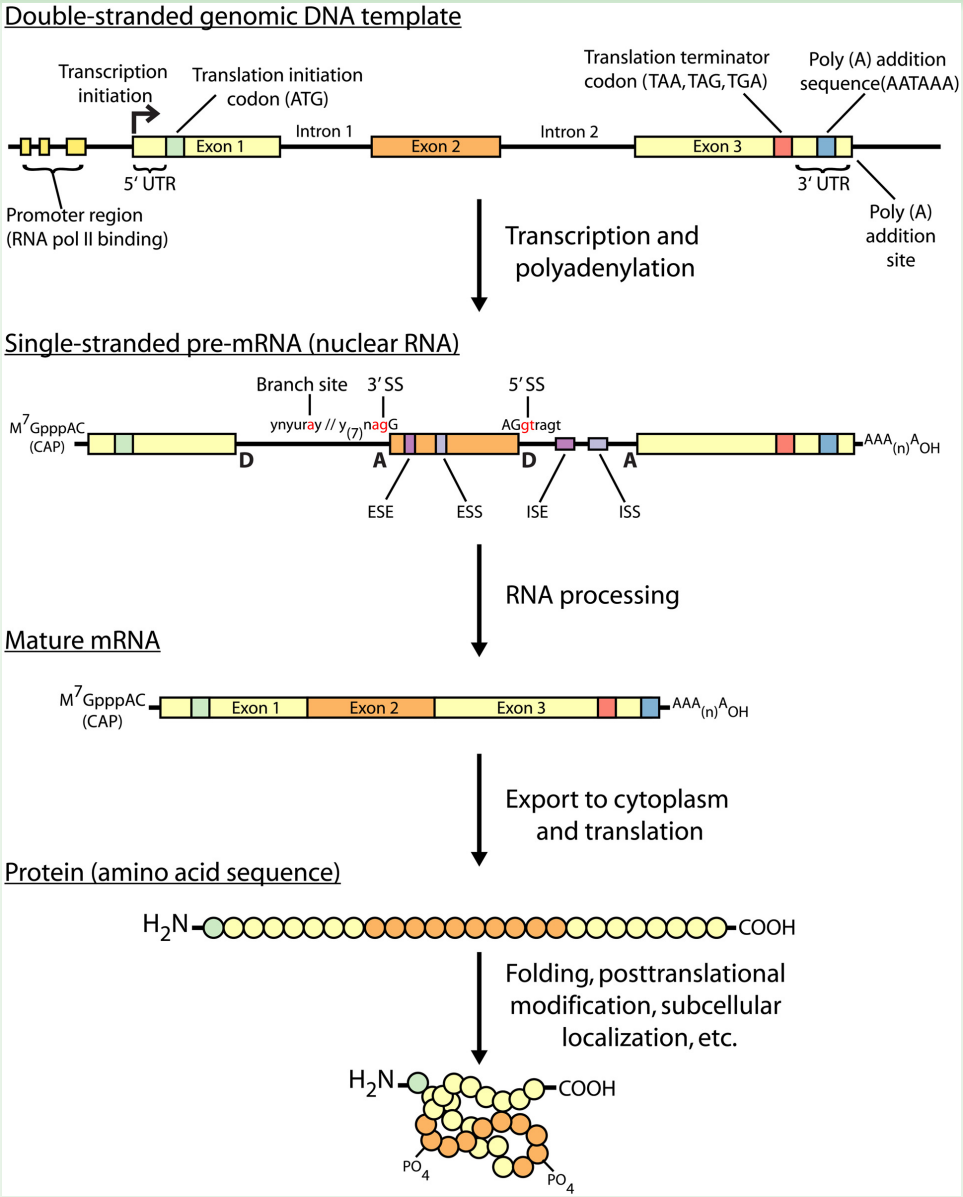
1. overview
2. Background: why RNA?
3. Ecological transcriptomics
4. RNA-seq workflow
5. Reference and *denovo* RNA-seq studies

Why RNA?

Why should I do RNA-seq, I have a genome?

- Insights into functional responses
- Genome = potential gene function
- RNA is better proxy for proteins than genome
- Genomes overemphasise mutations
  - Alternative splicing of isoforms, fusion transcripts
  - RNA interference genes may be silenced
  - Same genome in all cells and tissue types e.g brain cells and liver cells
  - Gene models for the identification of transcripts remains a challenge
- changes in experimental conditions = changes in gene expression
- Expression profile is a fingerprint of cell or tissue type
- Some genes are driven by gene-gene interations e.g epistasis

DNA makes RNA, and RNA makes protein

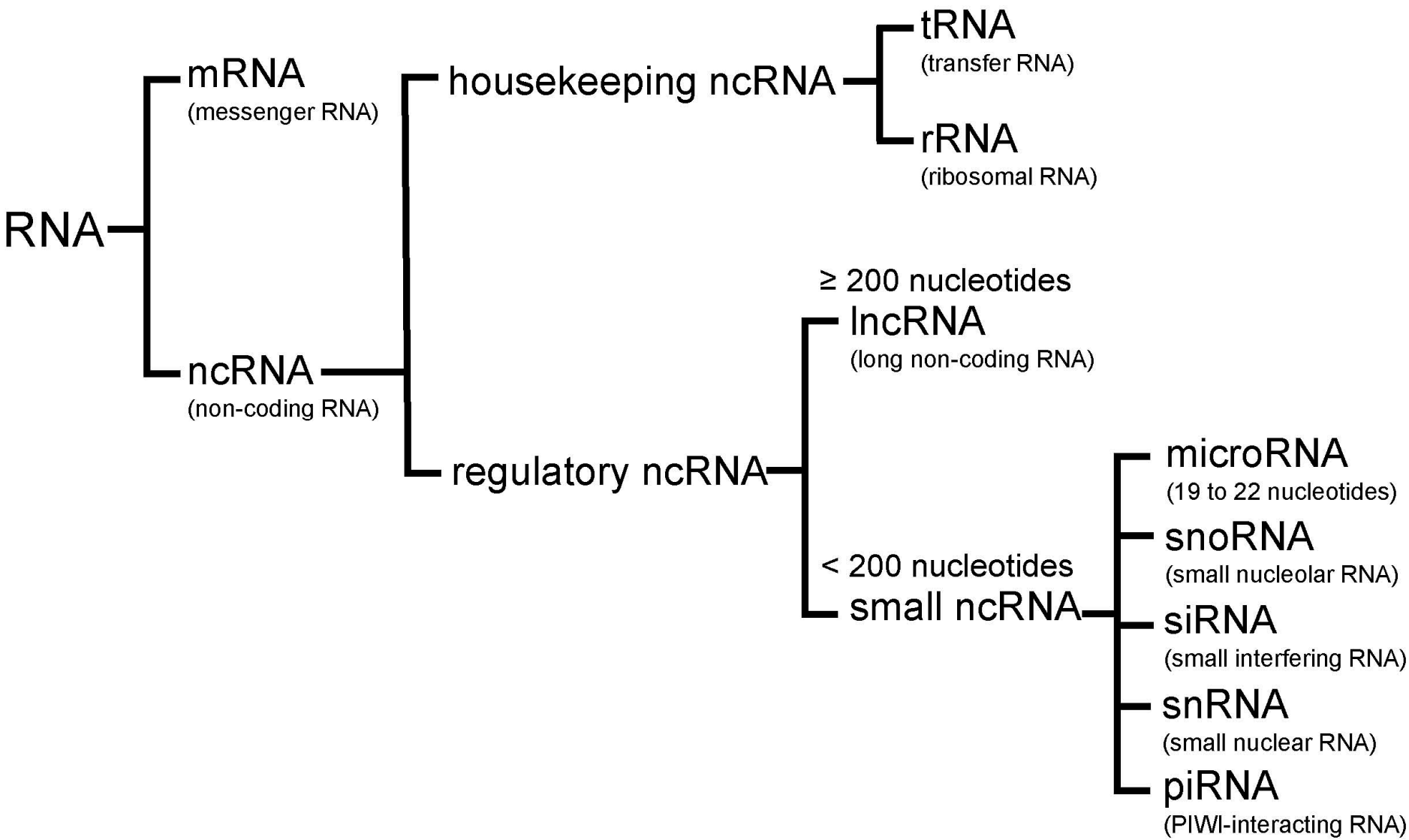


- We want to study proteins but they are hard to identify
- So we study RNA as a proxy for proteins

Abbreviations:
D=donor splice site;
A=acceptor splice site;
poly (A)=polyadenylation;
UTR= untranslated region;
SS=splice site;
ESE=exonic splicing enhancer;
ESS=exonic splicing silencer;
ISE=intronic splicing enhancer
ISS=intronic splicing silencer

10.1371/journal.pcbi.1004393

There are many RNA types



Ecological transcriptomics

Transcriptome

"The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition"

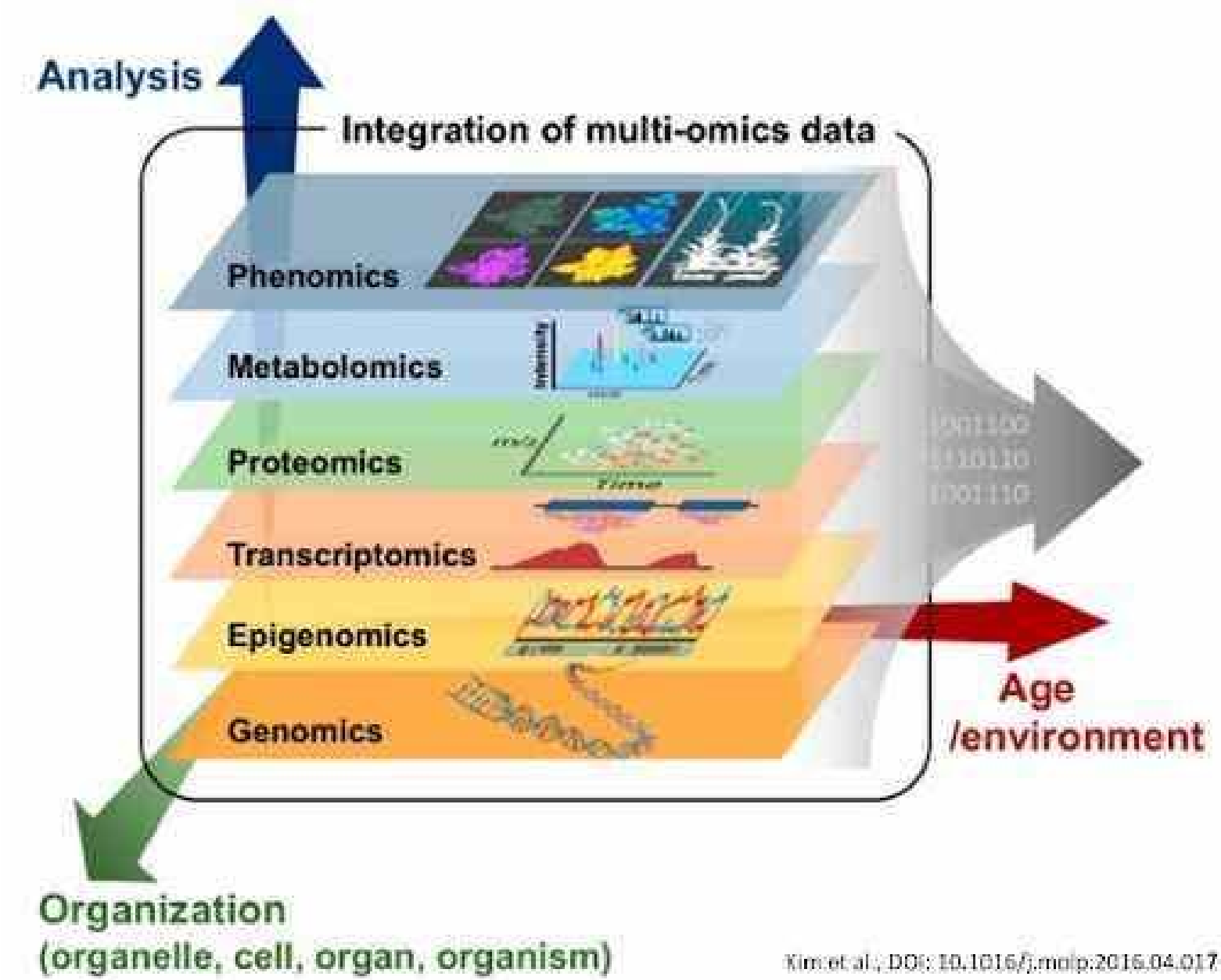
wang˙rna-seq:˙2009

Ecological transcriptomics?

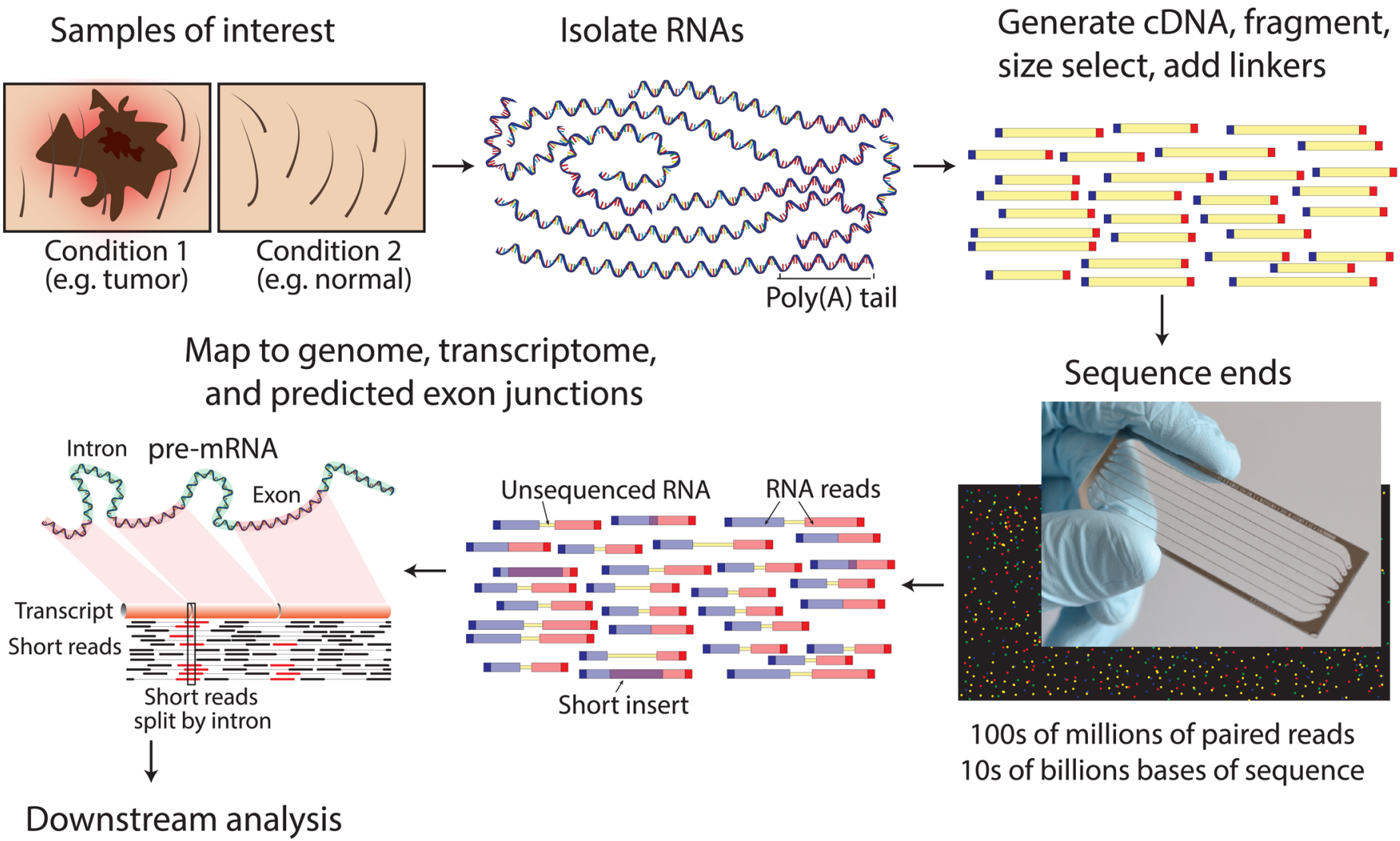
- Integration of transcriptomic analysis with ecological studies.
- Elucidation the genomic basis of phenotypic variation of individuals in response to environmental changes under natural or laboratory conditions

doi:https://doi.org/10.1002/9781444312249.ch12; **alvarez˙ten˙2015**

Omics layers:

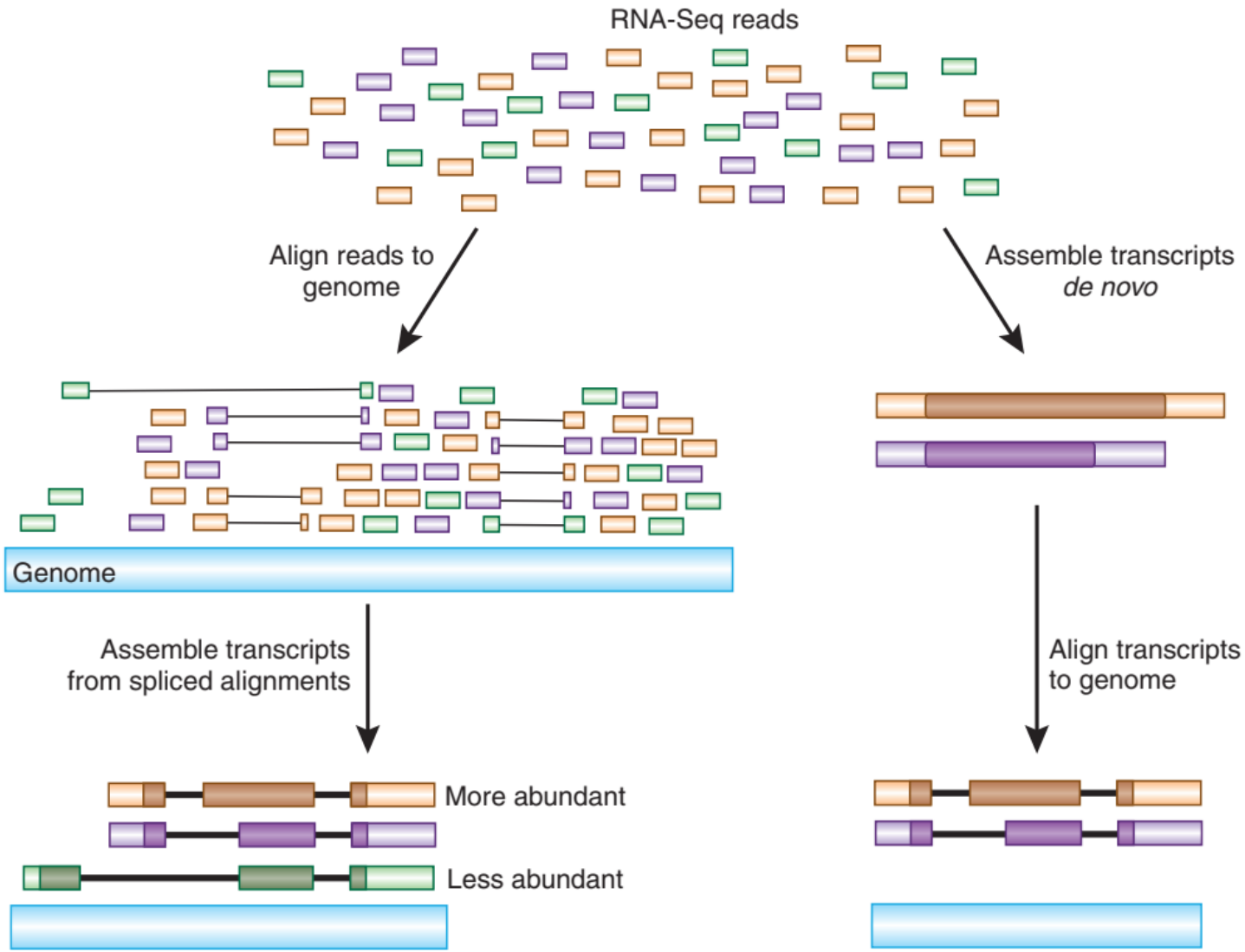


Overview of RNA-seq workflow



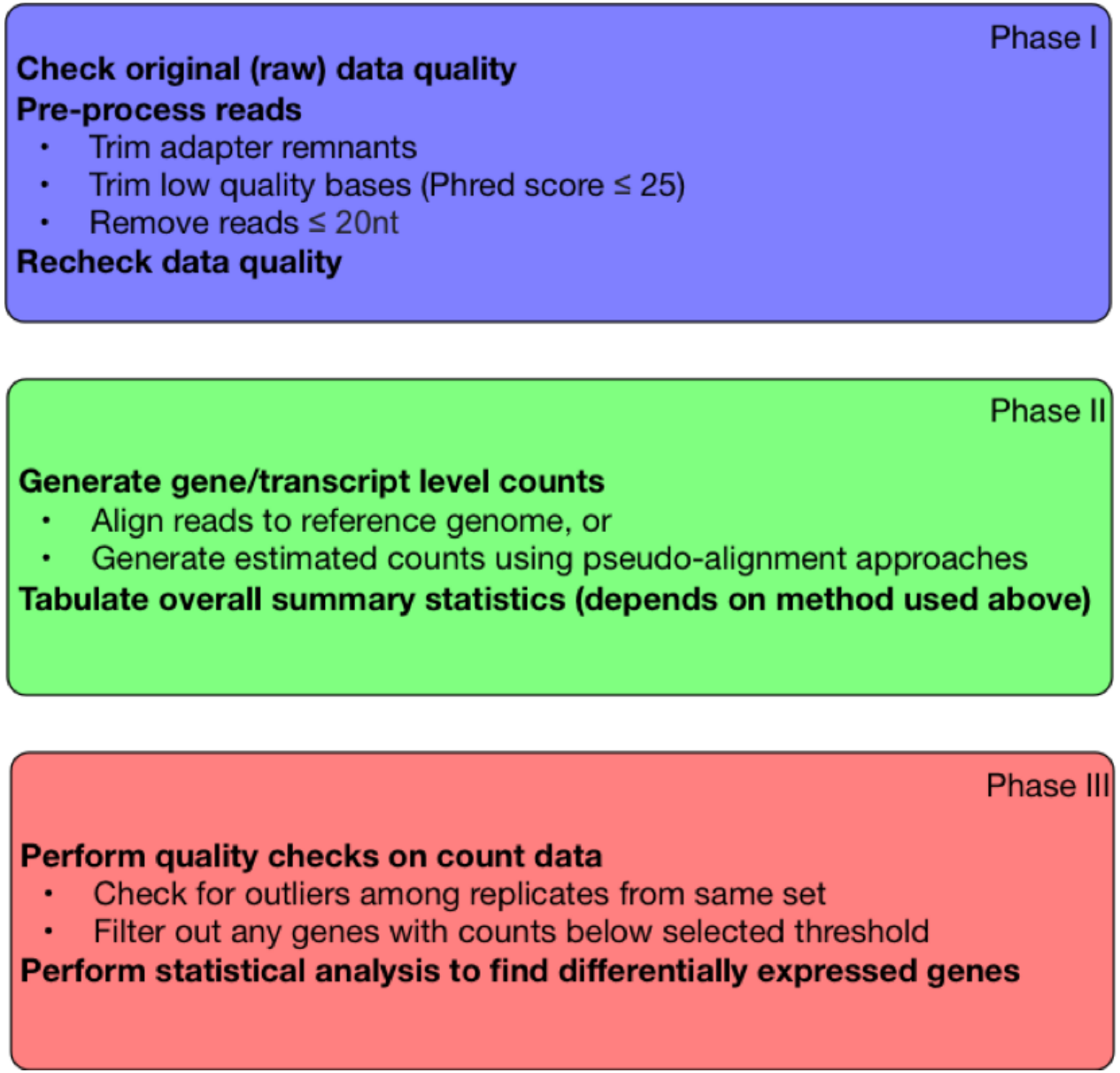
10.1371/journal.pcbi.1004393

Reference and *denovo* RNA-seq studies



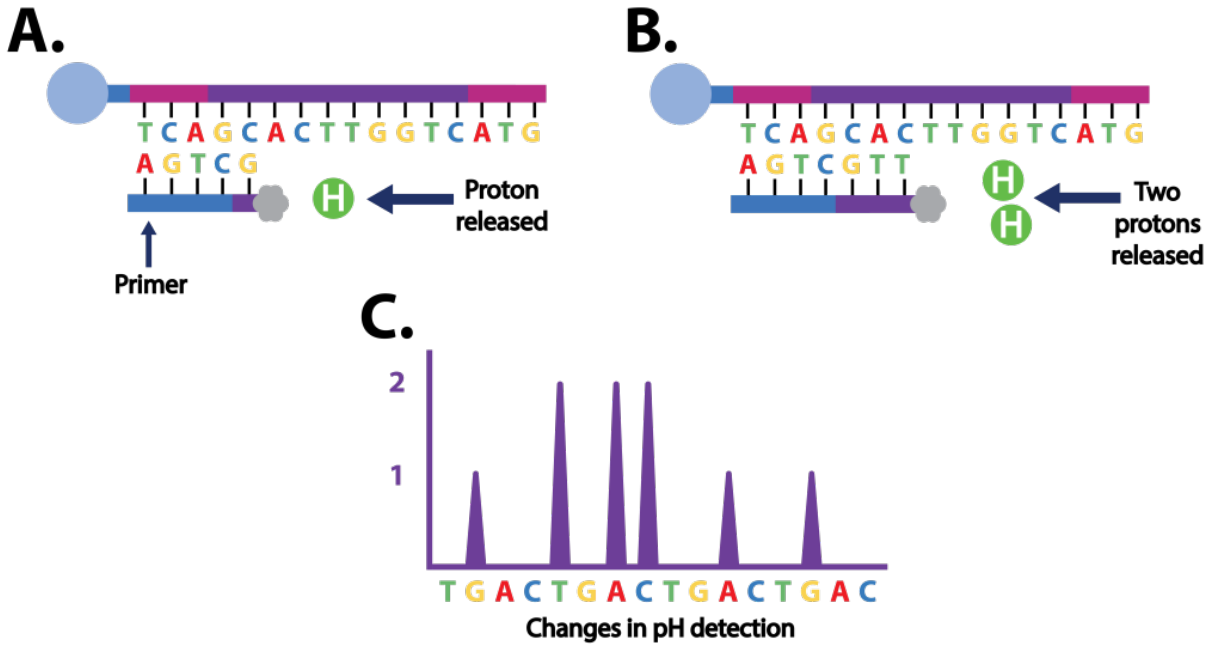
haas˙advancing˙2010

Overview of plan of workflow development



https://h3abionet.github.io/H3ABionet-SOPs/RNA-Seq-1-2.html

Ion torrent (Ion semiconductor sequencing)



https://apollo-institute.org/ion-torrent-sequencing/

Comparison of specifications of Illumina and Ion Torrent platforms.

Platform	Illumina MiSeq	Ion S5/Ion S5 Plus/Ion S5 Prime
Sequence yield per run	7.5–8.5 Gb on reagents v.2 12.5–15 Gb on reagents v.3	1,2-2 Gb on 520 chip 6-8 Gb on 530 chip
Accuracy	70% > Q30 at 600 cycles, 85% > Q30 at 500 cycles	85% > Q20
Systematic error	substitutions in GGC and GGT context	indels in homopolymer regions
Read length	500 (250 + 250) bp for reagents v.2 600 (300 + 300) bp for reagents v.3	~400 bp for double chip up to ~600 bp for single chip
Run Time	39 h for 500 cycles 56 h for 600 cycles (includes cluster generation, sequencing and base calling)	19,5 h for 400 bp (includes presequencing chip processing, initialization a sequencing)
Paired reads	Yes	No
Insert size	up to 550 bp	400 bp (up to 600 bp)

Step 1.1: Quality check

Example fastq file

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:N:18:1
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

Sequence reads: FASTQ Format Specification

- First line is the sequence header which starts with an '@' (not a '>')
- Everything from the leading '@' to the first whitespace character is considered the sequence identifier
- Everything after the first space is considered the sequence description
- Second line is the sequence
- Third line starts with '+' and can have the same sequence identifier appended (but usually doesn't anymore)
- Fourth line are the quality scores

`https://en.wikipedia.org/wiki/FASTQ_format`
`https://learn.gencore.bio.nyu.edu/ngs-file-formats/fastq-format/`

Phred quality scores

```
+SEQ_ID
!''*(((((***+))%%%++)(%%%%).1***
```

A quality value Q is an integer representation of the probability p that the corresponding base call is incorrect.

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

FASTQC: A quality control tool for high throughput sequence data

https://h3abionet.github.io/H3ABionet-SOPs/RNA-Seq-2-1.html

Before we start: an brief interlude on installing your tools

I highly recommend learning how to use conda

`https://conda.io/projects/conda/en/latest/index.html`

Running FastQC

```
fastqc seqfile1 seqfile2 .. seqfileN

fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam]
       [-c contaminant file] seqfile1 .. seqfileN
```

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

MultiQC

Aggregate results from bioinformatics analyses across many samples into a single report

`https://multiqc.info/modules/fastqc/`